Procesamiento del Lenguaje Natural, Revista nº 63, septiembre de 2019, pp. 111-118 recibido 29-03-2019 revisado 08-05-2019 aceptado 15-05-2019

Applying Siamese Hierarchical Attention Neural Networks for multi-document summarization

Aplicando Redes Neuronales Siamesas con Atención Jerárquica para resúmenes multi-documento

José Angel González, Julien Delonca, Emilio Sanchis, Fernando García-Granada, Encarna Segarra

VRAIN Valencian Research Institute for Artificial Intelligence
Universitat Politècnica València
Camino de Vera s/n
46022 Valencia
{jogonba2, esanchis, fgarcia, esegarra}@dsic.upv.es

Abstract: In this paper, we present an approach to multi-document summarization based on Siamese Hierarchical Attention Neural Networks. The attention mechanism of Hierarchical Attention Networks, provides a score to each sentence in function of its relevance in the classification process. For the summarization process, only the scores of sentences are used to rank them and select the most salient sentences. In this work we explore the adaptability of this model to the problem of multi-document summarization (typically very long documents where the straightforward application of neural networks tends to fail). The experiments were carried out using the CNN/DailyMail as training corpus, and the DUC-2007 as test corpus. Despite the difference between training set (CNN/DailyMail) and test set (DUC-2007) characteristics, the results show the adequacy of this approach to multi-document summarization.
Keywords: Siamese Hierarchical Attention Neural Networks, multi-document summarization

Resumen: En este artículo presentamos una aproximación al problema de resumen automático multi-documento, basada en Redes Siamesas Jerárquico-Atencionales. El mecanismo de atención de las redes Jerárquico-Atencionales permite asignar un peso a cada frase en función de su relevancia en el proceso de clasificación. Durante la generación del resumen sólo se tienen en cuenta los pesos asociados a las frases para seleccionar aquellas más relevantes. En este trabajo exploramos la posibilidad de adaptar estos modelos al problema de resumen multi-documento (típicamente documentos muy largos donde la aplicación directa de redes neuronales no se comporta correctamente). Se ha experimentado utlizando el corpus CNN/DailyMail para entrenamiento, y el corpus DUC-2007 para evaluación. A pesar de la heterogeneidad de las características entre el corpus de entrenamiento (CNN/DailyMail) y el corpus de test (DUC-2007), los resultados muestran la adecuación de esta propuesta al resumen multi-documento.
Palabras clave: Redes Neuronales Siamesas Jerárquico-Atencionales, resúmenes multi-documento

1 Introduction

Nowadays, the development of automatic summarization systems is an important challenge, due to the necessity of tackling with the large amount of information that is accessible in the web or in other repositories. There are many applications that could be enriched with summarization systems, such as news and tourist information websites, seminars or conference abstracts, etc.

Although there are some attempts to address the problem of audio and video summarization, the main efforts until now have been done for developing systems that consider text documents as input. Different strate-

ISSN 1135-5948. DOI 10.26342/2019-63-12

gies to the summarization problem have been proposed (Lloret and Palomar, 2012) (Tur and De Mori, 2011) (Saggion and Poibeau, 2013). It must be distinguished among abstractive summarization, where the summary is composed by sentences that does not appear in the document but contain almost all the meaning; extractive summarization, where the summary consist of a selection of the more salient sentences of the document; and mixed summarization, where summaries are generated by combining abstractive and extractive methods (See, Liu, and Manning, 2017). Due to the difficulty of developing good abstractive and mixed strategies, most of the approaches are extractive. These approaches are a good solution in some tasks, such as summarization of news, because the journalistic writing style tends to contain the main information in some few sentences, that usually appear at the beginning of the article.

Related to methodologies, due to the difficulty of obtaining training corpus of document-summary pairs to train supervised systems, most of the initial works were based on unsupervised techniques. This is the case of the statistical word features extraction (Carbonell and Goldstein, 1998), the obtention of latent concepts by means of Latent Semantic Analysis (Deerwester et al., 1990), the graph based approaches such as LexRank (Erkan and Radev, 2004), among others (Tur and De Mori, 2011)(Lloret and Palomar, 2012). On the other hand, some systems based on supervised techniques were proposed when manually training corpus were built. This is the case of summarization based on Support Vector Machines (Begum, Fattah, and Ren, 2009) or Conditional random Fields (Shen et al., 2007).

In order to promote the comparison of different summarization techniques, some conferences were organized. Two of the most important were DUC[1] and TAC[2] conferences, where the corpus used for evaluation consists of news obtained from different press agencies. The summaries are provided by humans in both cases.

Given the success of deep learning methods for Neural Networks (NN) in many applications of language technologies, some attempts to apply these techniques to document summarization have been done (Cheng and Lapata, 2016) (Nallapati et al., 2016) (Nallapati, Zhai, and Zhou, 2017) (See, Liu, and Manning, 2017) (Paulus, Xiong, and Socher, 2017) (Narayan, Cohen, and Lapata, 2018). One of the problems for estimating accurate NN-based models is the availability of large and high-quality corpora. An important resource in this field is the CNN/DailyMail summarization corpus (Cheng and Lapata, 2016)(Nallapati, Zhai, and Zhou, 2017). It consists of news from CNN and DailyMail, and is composed of 312,084 document-summary pairs. Other corpora, as NewsRoom have been recently created (Grusky, Naaman, and Artzi, 2018). NewsRoom's summaries were written by authors and editors in the newsroom of news, sports, entertainment, financial, and other publications. To create the dataset, the NewsRoom's authors performed a Web-scale crawling of over 100 million pages from a set of online publishers.

In this paper, we present an approach to multi-document summarization based on Siamese Hierarchical Attention Networks (SHA-NN). One advantage of this kind of models is that they can learn from positive and negative samples, that in our case are document-summary pairs. A positive sample is a document and its corresponding summary, and a negative sample is a document and a summary of other document randomly chosen. This way, the model is trained as a binary classifier and it doesn't need any kind of apriori assignation of scores to sentences as it is the case of other NN-based summarization systems (Nallapati, Zhai, and Zhou, 2017).

For training purposes, the input of the Siamese network consists of document-summary pairs along with the information about if it is a positive or negative sample. The document is processed by a subnetwork and the summary is processed by the other subnetwork of the SHA-NN system. Furthermore, it has an attention mechanism that can be used to provide an score to each sentence of the input document. For test purposes, only a document is provided, that is, only a subnetwork is used, and the output is a weight associated to each sentence of the input document. This way, the summary is generated by a selection mechanism applied on the weighted sentences. Additionally, the training process converge in few hours, dif-

[1] https://duc.nist.gov/

[2] https://tac.nist.gov/tac

ferently from other NN-based systems that converge after several days. In a preliminary research we applied our system to the CNN/DailyMail corpus for single document summarization (González et al., 2019). The obtained results are in line with the state-of-the-art.

A limitation of NN-based approaches is that input documents must not be very long. This is due to the fact that current models have not enough capability to capture long term dependencies. Moreover, generally, there are some space and time constraints. Therefore, NN-based approaches can work reasonably well with short documents, as news of journals, but it is necessary to adapt them when longer documents must be summarized. This is the case of DUC-2007 summarization task, where each multi-document is composed by the addition of different short documents related to a topic. In order to address this problem with the SHA-NN system, in this work we have proposed an iterative process that successively provides the most salient sentences from shorter pieces of the multi-document until the 250 words length summary (required by DUC competition) is obtained.

2 Related work

The use of Deep Neural Networks have made substantial progress in many language technologies, such as extractive document summarization. Some initial works, such as (Cheng and Lapata, 2016), addressed the summarization process as a sequential binary classification problem where sentences are classified as candidates to be extracted or not. This is done by an encoder-decoder system enriched with an attention mechanism that is used to score the sentences. The sentences are encoded by Convolutional Neural Networks and Recurrent Neural Networks are used to score them, taking into account the encoded representations, and the previous labeled sentences. In (Nallapati, Zhai, and Zhou, 2017) the sentence selection is addressed as a sequence classification problem by using Hierarchical Attention Networks (Yang et al., 2016) but modeling more features than in the work of (Cheng and Lapata, 2016). Both works first assign a score to each candidate sentence, and then extract the most salient sentences.

Other recent work presents some variants of NN-based approaches to enrich the systems. This is the case of Wu and Hu, (2018) and Narayan, Cohen, and Lapata, (2018) were a reinforcement learning algorithm is applied, considering ROUGE Lin, (2004) measures as reward. In Zhou et al., (2018) a system where the sentence scoring and the selection mechanism are jointly learnt is presented. At each step during extraction, the sentence extractor reads the representation of the last extracted sentence, and uses it to score the relevance of the remaining sentences. Finally, in Al-Sabahi, Zuping, and Nadher, (2018) an attempt to take into ac-count the structure of the document as in-formation to be considered in the selection of sentences is presented. The model com-putes the score of each sentence by modeling several features as: content richness, salience with respect the document, redundancy re-spect the summary and the position in the document.

3 Corpus description

Since there are no large enough corpora to train complex supervised systems for multi-document summarization, we used the CNN/DailyMail corpus for training the SHA-NN system in order to evaluate it in a multi-document summarization task. The corpus was built from the journals news and the associated summary, consisting in some highlights manually done by journalists. It consist of 312,084 document-summary pairs and three sets were defined from it: a training set of 287,226 pairs, a development set of 13,368 pairs, and a test set of 11,490 pairs. The mean compression ratio is 14:1, i.e. the reference summaries have, in average, approximately 14 times less words than the articles.

The DUC-2007 corpus consists of a collection of newswire documents. Documents were organized in 45 topics, and each topic is composed by 25 documents. The summarization problem consist of obtaining a summary of 250 words for each topic. The average number of words for document is 11,927 and the mean compresion ratio is 50:1. The gold standard summaries were done by human experts and there are 4 summaries for each one of the 45 topics.

Some statistics of both corpora are shown in Table 1. It can be seen that the lengths of the articles and summaries are extremely high for DUC-2007 in compari-

	Sents	Words
CNN/DM Articles	28.2	765.4
CNN/DM Summaries	3.8	53.4
DUC-2007 Articles	1028.0	12065.0
DUC-2007 Summaries	13.1	244.0

Table 1: Average number of sentences and words of CNN/DailyMail and DUC-2007 corpora for articles and summaries

son to CNN/DailyMail. Concretely, the articles of DUC-2007 have 36 times more sentences than the articles in CNN/DailyMail. Something similar happens with the number of words, where DUC-2007 articles have 15 times more words than their counterparts in CNN/DailyMail. It seems that the sentences in the CNN/DailyMail are twice as long as in the DUC-2007. However, regarding to the summaries, although in DUC-2007 they have more sentences and words than the summaries in CNN/DailyMail, the proportionality between the lengths is lower than for the articles.

4 Siamese Hierarchical Attention Networks

The SHA-NN architecture is shown in Figure 1. The left subnetwork represents the model for the document, and right subnetwork is the model for the summary. Both subnetworks are Hierarchical Attention Networks (Yang et al., 2016) composed by Bidirectional Long Short Term Memory (BLSTM) (Hochreiter and Schmidhuber, 1997). For training purposes, the input of the Siamese network consists of both, the sequence of words $\boldsymbol{x} = \{\boldsymbol{w_{11}}, ..., \boldsymbol{w_{1W}}, ..., \boldsymbol{w_{T1}}, ..., \boldsymbol{w_{TW}}\}$ of the document, and the sequence of words $\boldsymbol{x'} = \{\boldsymbol{v_{11}}, ..., \boldsymbol{v_{1V}}, ..., \boldsymbol{v_{Q1}}, ..., \boldsymbol{v_{QV}}\}$ of the summary, as well as the information about if it is a positive or negative sample, that is coded as 0 or 1 in the output $\boldsymbol{y}$. The input words are coded as d-dimensional embeddings, that are estimated with a skip-gram model from the CNN/DailyMail corpus. The output of the word level are the vector representation of sentences (in Figure 1, $\boldsymbol{s}_i$ for the document, and $\boldsymbol{q}_i$ for the summary), and the output of the sentence level are the vector representations of the document $\boldsymbol{r}$ and the summary $\boldsymbol{p}$. The boxes labeled as α_i and β_i represent the attention mechanism that assigns a score to each word or a sentence in the document side and the summary side.

Finally, the vector representation of the document $\boldsymbol{r}$ and the summary $\boldsymbol{p}$ as well as the difference between them ($|\boldsymbol{r}-\boldsymbol{p}|$) are concatenated in an output layer with a softmax activation function that works as classifier, as shown in Eq 1.

$$\widehat{y} = softmax(W_{\widehat{y}}[\boldsymbol{p}, \boldsymbol{r}, |\boldsymbol{r} - \boldsymbol{p}|] + b_{\widehat{y}}) \quad (1)$$

When the system is working on summarization mode, only the left side of the full model is considered (the subnetwork that processes the input document). A forward pass is performed on it to obtain the attention output, α, associated to each sentence of the input document, which allows us to generate a ranking of the most salient sentences to build summaries. Although many approaches can be used to select the most relevant sentences, in this work we have chosen the sentences considering only these attention outputs.

5 Multi-Document summarization process

The straightforward application of NN-based models to multi-document summarization on extremely long documents has several drawbacks. First, this kind of models have not got enough capacity to capture long term dependencies on extremely long sequences. Moreover, the longer these documents, the more complex these dependencies are and then, models must be more complex to explain these dependencies. That is, more parameters have to be estimated and therefore, more training data are required.

In addition, generally, there are space and time constraints. The most known and useful strategy in order to train NN-based models efficiently by using mini-batch training mode, consists of truncating the input document to a limited length of words and sentences. However, by doing this, some fragments of the input are discarded without a relevance criterion and some of these fragments could be relevant for computing the output. This is specially important in multi-document summarization where documents are compositions of many short single documents.

In order to address this problem, we have developed an iterative mechanism that first

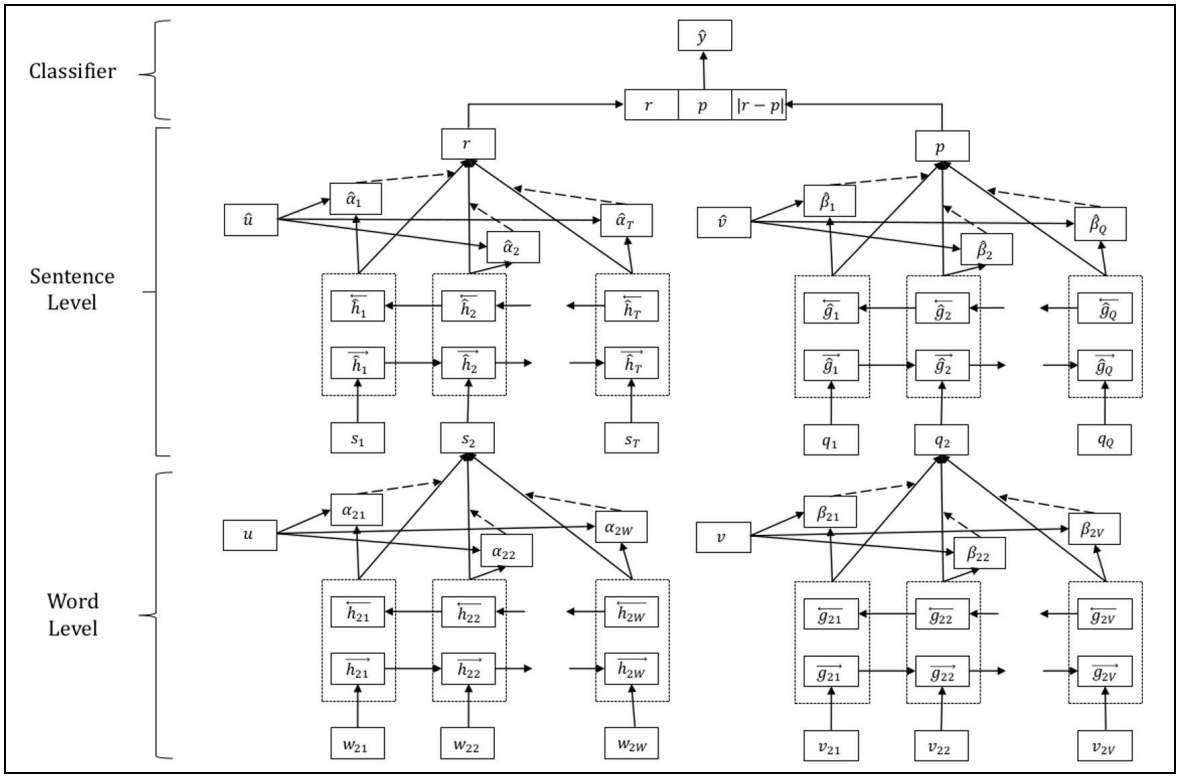

Figure 1: SHA-NN architecture proposed in (González et al., 2019)

obtains summaries from short fragments of the whole multi-document, and successively new summaries are generated from these previous summaries. Concretely, in each iteration of this iterative process, each fragment is separately summarized and then a new document is created for the new iteration by concatenating these summaries. This process is repeated until a summary of the desired length is obtained. The process is shown in Figure 2. Furthermore, this approach can be specially appropriate for the DUC corpus, because each long document is composed by the addition of shorter documents.

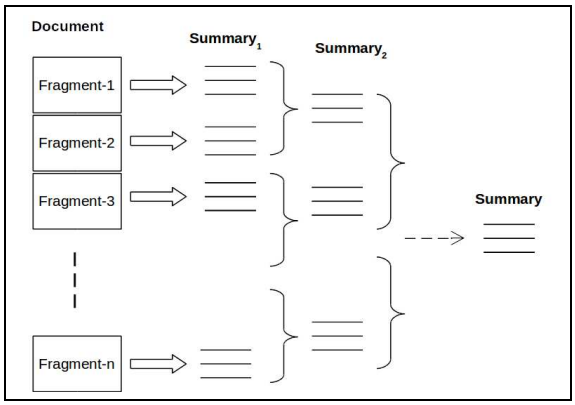

Figure 2: Iterative mechanism to obtain summaries from short fragments of the whole multi-document

6 Experiments

Some experiments were performed with the DUC-2007 corpus. Results were evaluated in terms of some ROUGE measures (Lin, 2004). In order to compare with other systems, we used the evaluation software given by the organizers of the competition, as well as the gold standard summaries also provided by them.

As the DUC-2007 corpus was designed only for evaluation, there is not possible to learn models with it. For this reason we have trained SHA-NN with the CNN/DailyMail corpus, and we have studied if a system trained with a type of corpus can generalize features that can be successfully applied for other types of corpora. One of the main differences between training and test corpus is the length of the input documents, that is much longer in the test. Another characteristic of the training corpus is that the named entities are anonimized, diferently than in the DUC corpus.

The summarization experiments were done as follows. For each topic, a first summary of each document belonging to this

topic is performed. To do this, each document is splitted in fragments of 20 sentences, and for each block, the 3 most salient sentences are extracted. The ratio between the number of input sentences of each fragment and the number of selected sentences for the summary, 20:3, is similar to the ratio considered in the training process with the CNN/DailyMail corpus. Once the multi-document associated to a topic is summarized, the obtained summaries are concatenated, and a new process of summarization is performed on that set of sentences in the same way. That is, the new set of sentences is splitted in blocks of 20 sentences and for each block the three most salient sentences are extracted. This process is done iteratively until arriving to 250 words, the length established by the competition. In the case of the experiments with the DUC-2007 corpus only two iterations have been performed, because after these two iterations the 250 word summaries were obtained. All this process is done for each one of the 45 topics.

Table 2 shows the results obtained in terms of Precision (P), Recall (R), and F-measure (F) for each one of the ROUGE measures. The results obtained are in line with those published in the DUC-2007 competition.

We did another experiment that uses a more simple way for selecting the summary sentences. In this experiment only one iteration of the summarization process is performed. This way, a set of sentences containing the three sentences of each fragment, as well as the associated score assigned by the SHA-NN system to each sentence, are obtained. These sentences are ranked considering their score, and they are sequentially selected until arriving to 250 words. Table 3 shows that results are slightly lower than those of the previous experiment. It seems that the iterative process takes advantage of the context generated by the salient sentences, instead of only consider the original context where relevant and not-relevant sentences participate.

	P	R	F
ROUGE-1	0.37098	0.37557	0.37204
ROUGE-2	0.07122	0.07240	0.07158
ROUGE-3	0.02209	0.02253	0.02225
ROUGE-4	0.01057	0.01074	0.01063
ROUGE-L	0.34084	0.34520	0.34190
ROUGE-W-1.2	0.18082	0.09862	0.12719
ROUGE-SU4	0.12767	0.12956	0.12819

Table 2: Results of the full iterative summarization process

	P	R	F
ROUGE-1	0.36946	0.37359	0.37113
ROUGE-2	0.06959	0.07028	0.06986
ROUGE-3	0.01958	0.01976	0.01965
ROUGE-4	0.00790	0.00800	0.00794
ROUGE-L	0.34123	0.34497	0.34274
ROUGE-W-1.2	0.18070	0.09844	0.12728
ROUGE-SU4	0.12366	0.12496	0.12418

Table 3: Results with only one iteration of the summarization process

7 Conclusions

We have presented in this work an approach to adapt a document summarization system based on Siamese Hierarchical Attention Neural Networks to a multi-document summarization task, the DUC-2007 competition. In the absence of an adequate enough large training corpus for this domain, the SHA-NN system has been trained on the CNN/DailyMail corpus, that presents some structural differences compared to the DUC-2007 corpus. It has been necessary to define a specific mechanism to allow the SHA-NN system to be applied to that multi-document summarization task. The results obtained with the DUC-2007 corpus are in line with those published in the DUC-2007 competition. As future works we will study if different reordering on the sentences obtained at each iteration can improve the results. It can be also interesting to use more information than just the score assigned to each sentence for selecting the most salient sentences. For example, in order to avoid including similar sentences in the summary, some distance among the candidate sentences can be taken into account.

Acknowledgements

This work has been partially supported by the Spanish MINECO and FEDER founds under project AMIC (TIN2017-85854-C4-2-R). Work of José-Ángel González is also financed by Universitat Politècnica de València under grant PAID-01-17.

References

Al-Sabahi, K., Z. Zuping, and M. Nadher. 2018. A hierarchical structured self-attentive model for extractive document summarization (hssas). *IEEE Access*, 6:24205–24212.

Begum, N., M. Fattah, and F. Ren. 2009. Automatic text summarization using support vector machine. 5:1987–1996, 07.

Carbonell, J. and J. Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 335–336.

Cheng, J. and M. Lapata. 2016. Neural summarization by extracting sentences and words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.

Deerwester, S., S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Erkan, G. and D. R. Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Int. Res.*, 22(1):457–479, December.

González, J.-Á., E. Segarra, F. García-Granada, E. Sanchis, and L.-F. Hurtado. 2019. Siamese hierarchical attention networks for extractive summarization. *Journal of Intelligent & Fuzzy Systems (JIFS)*, To be published.

Grusky, M., M. Naaman, and Y. Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 708–719, New Orleans, Louisiana, June. Association for Computational Linguistics.

Hochreiter, S. and J. Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In S. S. Marie-Francine Moens, editor, *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain, July. Association for Computational Linguistics.

Lloret, E. and M. Palomar. 2012. Text summarisation in progress: a literature review. *Artificial Intelligence Review*, 37(1):1–41.

Nallapati, R., F. Zhai, and B. Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA.*, pages 3075–3081.

Nallapati, R., B. Zhou, C. N. dos Santos, Ç. Gülçehre, and B. Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *CoNLL*, pages 280–290. ACL.

Narayan, S., S. B. Cohen, and M. Lapata. 2018. Ranking sentences for extractive summarization with reinforcement learning. In *Proceedings of NAACL-HLT*, pages 1747–1759. Association for Computational Linguistics.

Paulus, R., C. Xiong, and R. Socher. 2017. A deep reinforced model for abstractive summarization. *CoRR*, abs/1705.04304.

Saggion, H. and T. Poibeau, 2013. *Automatic Text Summarization: Past, Present and Future*, pages 3–21. Springer Berlin Heidelberg, Berlin, Heidelberg.

See, A., P. J. Liu, and C. D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083. Association for Computational Linguistics.

Shen, D., J.-T. Sun, H. Li, Q. Yang, and Z. Chen. 2007. Document summarization using conditional random fields. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, pages 2862–2867.

Tur, G. and R. De Mori. 2011. *Spoken language understanding: Systems for extracting semantic information from speech.* John Wiley & Sons.

Wu, Y. and B. Hu. 2018. Learning to extract coherent summary via deep reinforcement learning. In *Thirty-Second AAAI Conference on Artificial Intelligence.*

Yang, Z., D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489. Association for Computational Linguistics.

Zhou, Q., N. Yang, F. Wei, S. Huang, M. Zhou, and T. Zhao. 2018. Neural document summarization by jointly learning to score and select sentences. *CoRR*, abs/1807.02305.